http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7427-7434, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7427.7434

Residential Hedonic Price Multivariate Linear Regression Model Based on Approximative Normalization

Wang Heyong and Hong Ming
College of E-Business, South China University of Technology, Guangzhou, China

Abstract: Hedonic price model and multiple regression analysis are the commonly used methods in the study of residential price. It affects the accuracy of the residential hedonic price multivariate linear regression model that the distributions of residential characteristic variables do not meet the normal distribution. According to the problem, this paper presents the approximative normalization. Through empirical model, approximative normalization improves the residential hedonic price multivariate linear regression model in the aspect of the accuracy to predict reality and empirical data are used to confirm its effectiveness.

Key words: Approximative normalization, residential hedonic price multivariate linear regression model

INTRODUCTION

It is a major issue faced by housing developers to determine reasonable new real estate sales prices and accurately grasp the market changes to make timely modifications or adjustments, on the one hand to win customers as well as the market and to maximize profits on the other hand. The traditional methods to determine estate sales prices include cost-oriented method, market-oriented method and demand-oriented method. The most widely used method currently is to determine the estate sales price based on the market price using hedonic price model (Lu, 2012).

Hedonic price model is commonly used in residential price valuation in foreign countries. The model evaluate residential prices from the implicit prices of real estate features, to overcome various defects of traditional methods and can get good results. Various domestic and foreign researches have studied residential price valuation using hedonic price model. Domestically, Wang (2006) analyzes urban residential hedonic price model theoretically and Ma and Li (2003), Wen and Jia (2006), Hao and Chen (2007), Zhang and Chen (2008) and Guo *et al.* (2006) analyze the factors affecting residential prices in particular regions using hedonic price model and construct their particular hedonic price models. Abroad butler (1982), Ozanne and Malpezzi (1985) and Crocker *et al.* (1987) study the selection of factors that affect residential prices.

Multiple regression analysis is one kind of the analytical methods to study the relationship between residential price and factors that affect it. Ren *et al.* (2009), Liu *et al.* (2009) and Wang (2008) build the residential price model using multiple regression analysis. However, wrong estimations will exist and even the credibility of model predictions will be endangered when establishing residential price models because residential characteristic variables are too many and complex. This situation is more likely to occur when establishing residential price model using the multiple linear regression model when the residential characteristic variables are not normally distributed (Dunn and Clark, 1987).

Study in this paper is based on hedonic price model and multiple linear regression analysis. First, distributions of the dependent variable and the independent variables in the model are discussed according to real empirical data. Then approximative normalization is processed. This paper use "approximative normalization" to represent the process that the distributions of variables whose distribution do not meet the normal distribution are approximatively or completely transformed to normal distribution. After that, a model dealt with approximative normalization is built to compare with the original one not dealt with approximative normalization. Finally, this paper presents the result whether approximative normalization will optimize the residential hedonic price multivariate linear regression model.

SELECTION OF RESIDENTIAL CHARACTERISTIC VARIABLES

The following three types of residential features are commonly considered important in the domestic and foreign studies [3, 15]:

Corresponding Author: Wang Heyong, College of E-Business, South China University of Technology, Guangzhou, China

Table 1: Selection of residential characteristics

Type	Residential characteristics variables
BSF	total floors, floor, residential age, residential area, decoration level, orientation, No. of rooms, No. of living rooms, No. of bathrooms, No. of balconies, property type, packing spaces
RC	floor area rate, greening rate, public facilities, education facilities, sports and commercial facilities
NEC	area, transport facilities, CBD distance

- Building Structure Features (BSF), such as the housing area, orientation
- Regional Characteristics (RC), such as the distance from public transportation to the city center.
- Neighborhood Environmental Characteristics (NEC), such as the residential infrastructure, culture and entertainment

This study has selected 20 residential characteristic variables including 12 building structure features, 5 regional characteristics and 3 neighborhood environmental characteristics. The variables have been chosen are shown in Table 1.

QUANTIFICATION OF RESIDENTIAL CHARACTERISTICS VARIABLES

In the process of establishing residential hedonic price models, non-numerical data need to be quantified. Methods commonly used for Quantification include comprehensive index method, numerical method, dummy variables method, Likert scale method and fuzzy mathematics method. The non-numerical residential characteristics variables chosen in this paper are quantified using one the methods mentioned above (shown in Table 2) and the numerical variables retain their original values.

Regional characteristics: In Guangzhou City, economic development level and consumption of residential market in a particular area may differ from the others. Considering this situation, Likert scale method is chosen to quantify the following 13 areas in Guangzhou City (Shown in Table 3.

Comprehensive index method is chosen to quantify this transport facilities according to survey of residential traffic conditions in each area. The values of this variable are 0 at the beginning and they increase according to the following rules:

Within 1km away from the house:

- If there is any subway station, then the values increase by 1
- If there are 10 or more bus lines, then the values increase by 2
- If there are 5 to 9 bus lines, then the values increase by 1

Table 2: Quantification of residential characteristics variables

Type	Variables	Method used
BSF	Floor	Likert scale
	Decoration level	Likert scale
	Orientation	Likert scale
	Property type	Likert scale
	Packing spaces	Dummy variables
RC	Area	Likert scale
	Transport facilities	Comprehensive index
NEC	Public facilities	Likert scale
	Education facilities	Comprehensive index
	Sports and commercial facilities	Comprehensive index

Table 3: Area quantification

Area	Quantified value
Yuexiu	3
Tianhe	3
Haizhu	2
Panyu	2
Liwan	2
Baiyun	2
Huangpu	1
Zengcheng	1
Huadu	1
Conghua	1
Nansha	1
Luogang	1
Around Guangzhou city	1

Table 4: Transport facilities quantification

Within 1km away from house	Score
numOf (subway station)>= 1	+1
numOf (bus lines)>= 10	+2
9>numOf (bus lines)>= 5	+1
numOf (bus lines)<5	+0

- If there are less than 5 bus lines, then values increase by 0

The rules are summarized in Table 4. The function numOf (a) output the number of a.

For example, if the transport facilities of a house meets condition 1 and condition 2 in Table 3-3, then the final value of the variable is 3 (= 0+1+2). The greatest value of this variable is 3 and the smallest is 0. The higher score, the transport facilities are better.

Building structure features: Likert scale method is chosen to quantify floor according to the principle that the higher floor, the residential price is higher (except the highest floor). Let k = floor/total floors, the Quantification rules are shown in Table 5.

Likert scale method is chosen to quantify decoration level. The higher decoration level, the value is greater (shown in Table 6).

Table 5: Floor quantification

Conditions of k	Quantified value
k<1/3	0
2/3>k>= 1/3	1
k>= 2/3	2

Table 6: Decoration level quantification

Decoration level	Quantified value
Blank	1
Simple	2
Media	3
Top	4
Luxurious	5

Table 7: Orientation quantification

Orientation	Quantified value
East	1
West	1
South	2
North	2
East-West	2
South-East	3
North-East	3
North-West	3
South-West	3
North-South	4

Table 8: Property type quantification

Property type	Quantified value
Ordinary residential	1
apartment	2
villa	3

Table 9: Parking spaces quantification

Have parking spaces?	Quantified value
Yes	1
No	0

Table 10: Public facilities quantification

Level	Quantified value
Excellent (n>= 10)	3 or 4
Good (10>n>= 2)	1 or 2
Not Bad (n<2)	0

Likert scale method is chosen to quantify orientation according to traditional customer preferences in China. The better orientation, the greater value (shown in Table 7).

Likert scale method is chosen to quantify property type (shown in Table 8).

Dummy variables method is chosen to quantify parking spaces (shown in Table 9).

Neighborhood environmental characteristics: Likert scale method is chosen to quantify public facilities. Let n represents number of public facilities. The greater n, the value is greater. The result is shown in Table 10.

Comprehensive index method is chosen to quantify education facilities. The original values are 0 and they increase according to the following rules. Within 1km away from the house:

Table11: Education facilities quantification

Within 1km away from house	Score
Kindergartens	+1
Primary schools	+1
Secondary schools	+1
Universities	+1

Table 12: SCF quantification

Within 1km away from house	Score
Sports facilities	+1
Integrated shopping malls	+1
Banks	+1
Hospitals	+1
Postal services	+1
Leisure facilities	+1

- If there is any kindergarten, then the values increase by 1
- If there is any primary school, then the values increase by 1
- If there is any secondary school, then the values increase by 1
- If there is any university, then values increase by 1
- The result is shown in Table 11

Comprehensive index method is chosen to quantify sports and commercial facilities (SCF). The original values are 0 and they increase according to the following rules:

Within 1km away from the house:

- If there is any sports facility, then the values increase by 1
- If there is any integrated shopping mall, then the values increase by 1
- If there is any bank, then the values increase by 1
- If there is any hospital, then the values increase by 1
- If there is any postal service, then the values increase by 1
- If there is any leisure facilities, then the values increase by 1

The result is shown in Table 12.

MULTIPLE LINEAR REGRESSION ANALYSIS

Multiple linear regression model: Suppose there is a linear correlation between a random variable and m (m = 1, 2, 3, ...) non-random variables, the relationship between them can be expressed by the following multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \mu \quad (4\text{-}1)$$

In the model 4-1, y is the dependent variable and x_j (j = 1, 2, 3, ..., m) are the independent variables, β_j (j = 1, 2, 3, ..., m) are the partial correlation coefficients and μ is the random error.

Dunn and Clark (1987) pointed out that the traditional multiple linear regression analysis is applicable for normally distributed and complete data. However, it is found in the distribution test of the experimental data sets that nearly all the continuous numerical variables (including residential price) do not meet normal distribution. Theoretically, it affects the validity of the multiple linear regression model whether variables in the model follow normal distribution. Those variables not normally distributed can be approximatively or completely transformed to normal distribution (e.g. logarithmic), which this paper calls "approximative normalization".

EXPERIMENTS

Experimental preparation: The experimental data used in this paper are collected by two ways. Most of them such as building structure features variable are collected from professional real estate information websites such as gz.soufun.com. Data of some regional characteristics and some neighborhood environmental characteristics can not be collected directly and these are collected from Baidu Map and Guangzhou Electric Map.

Totally, 407 complete records are collected. These records are divided into training set with 70% of the records (274 records) and testing set with 30% of the records (133 records). In order to maintain consistency characteristic variables.

Analysis software used in the experiment is SPSS Clementine.

Experimental steps: The experimental premise is that the overall multiple linear regression model is significant (F-test only, no t-test) in order to keep the residential characteristics variables the same in different models. Let the significance level $\alpha = 0.005$, the experimental steps are following:

- Discussing the distribution of dependent variables as well as independent variables and Processing approximative normalization
- Establishing multiple linear regression models with original data set and then establishing another with data set dealt with by approximative normalization
- Comparing the two models established in step (2) to conclude that whether the latter is better than the former according to two indexes used in this study

The two indexes are R Square and correlation between actual values and predicted values.

The correlation is calculated by testing set.

Table 12: Functions for approximative normalization

Characteristics	Function
Residential price	lnx
Residential area	lnx
Floor area rate	$x^{1/2}$
Greening rate	lnx
CBD distance	lnx

Table 13: ANOVA1

Source	SS	df	MS	F
Regression	9779555.1	20	488977.8	30.1
Residual	4110145.9	253	16245.6	
Total	13889701.0	273		

Distributions of variables and Approximative Normalization: Among all the variables, residential price, residential area, floor area ratio, greening rate, CBD distance and are numerical continuous variables whose distributions can be transformed by approximative normalization. Functions used for approximative normalization include x^{-1}, lnx, $\log_{10} x$ and $x^{1/2}$.

The selection rule to choose the right transformations here is that the distribution of variables meet normal distribution completely or approximatively and values of variables are not too small or too big.

Figure 1 shows the distributions of residential price.

According to the selection rule, lnx is chosen for residential price variable.

Figure 2 shows the distributions of residential area.

According to the selection rule, lnx is chosen for residential area variable.

Figure 3 shows the distributions of floor area rate.

According to the selection rule, $x^{1/2}$ is chosen for floor area rate variable.

Figure 4 shows the distributions of greening rate.

According to the selection rule, lnx is chosen for floor area rate variable.

Figure 5 shows the distributions of CBD distance.

According to the selection rule, lnx is chosen for CBD distance variable.

Table 12 shows the functions chosen for approximative normalization.

Residential hedonic price multivariate linear regression model: RMR model is short for residential hedonic price multivariate linear regression model in this paper. Residential price variable is the independent.

First, a RMR model is established using original data set, the result is shown in Table 13.

From Table 13, $F_{0.005}(20.253) = 2.00 < 30.1$, that indicates the overall model is significant. This model is called original model in this paper.

Next, another RMR model is established using approximative normalized variables. The result is shown in Table 14.

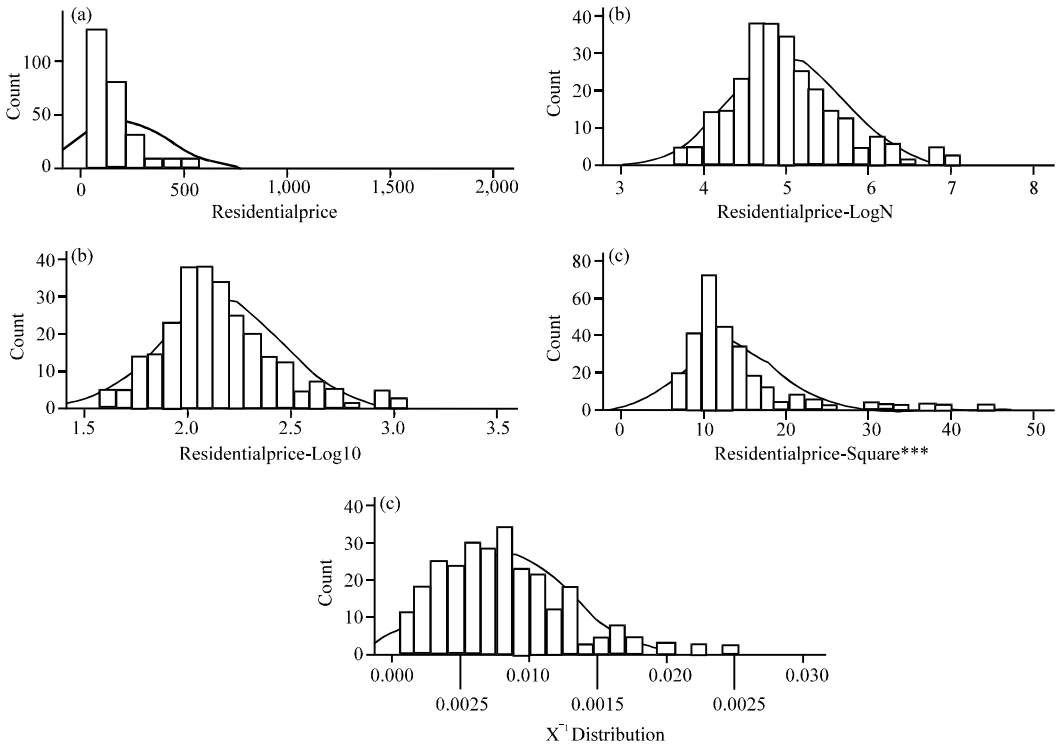

Fig. 1(a-e): Distributions of residential price (a) Current distribution, (b) lnx distribution, (c) log10×Distribution x1/2, (d) Distribution and (e) x 1 Distribution

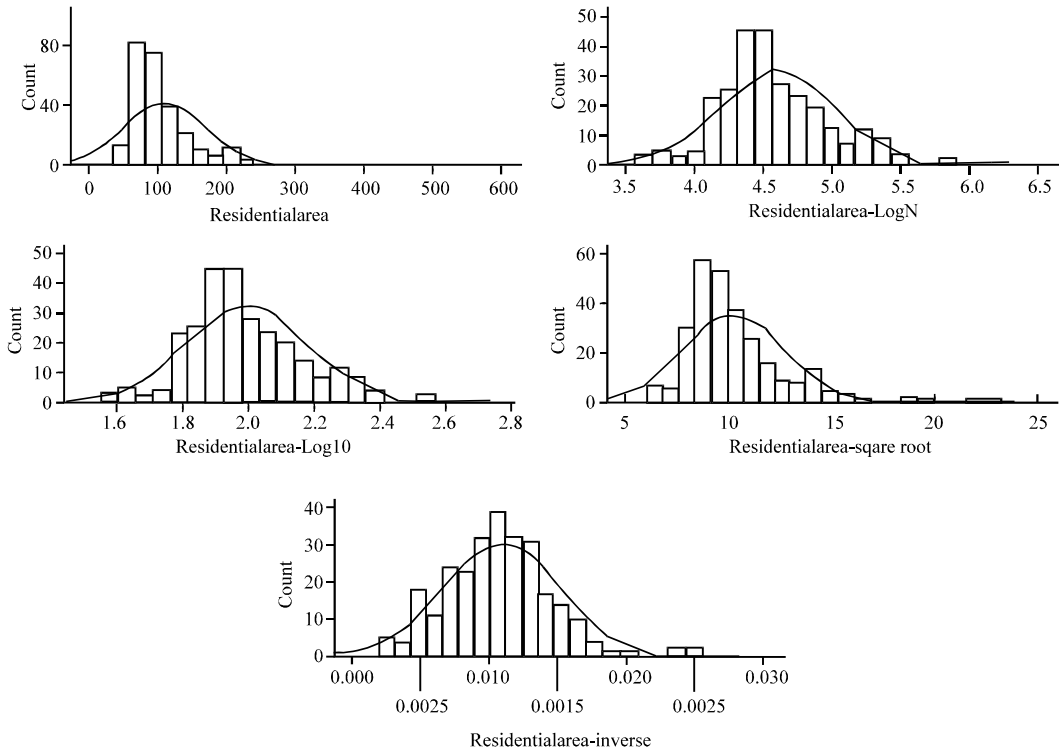

Fig. 2(a-e): Distributions of residential area (a) Current distribution, (b) lnx distribution, (c) log10×Distribution x1/2, (d) Distribution and (e) x 1 Distribution

From Table 14, 13, $F_{0.005}(20.253) = 2.00<57.864$, that indicates the overall model is significant. This model is called approximatively normalized model in this study.

Experimental Results: By comparing the two models established in section 5.4 using the three indexes mentioned in section 5.2, it is found that approximative normalization improves the original model.

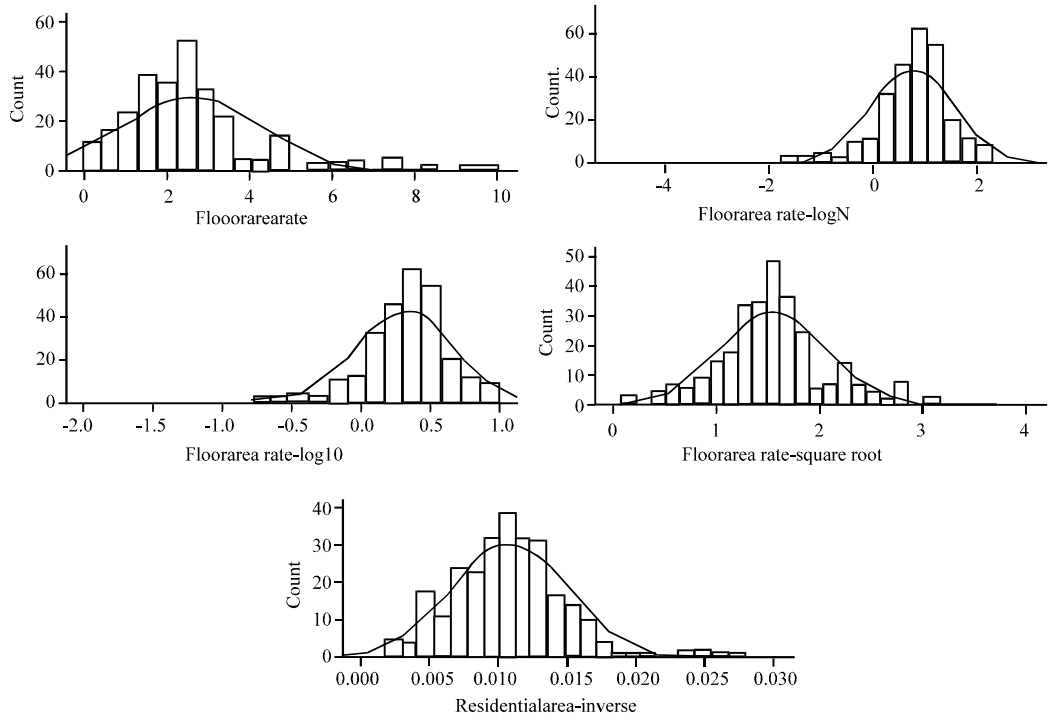

Fig. 3(a-e): Distributions of floor area rate (a) Current distribution, (b) lnx distribution, (c) log10×Distribution x1/2, (d) Distribution and (e) x 1 Distribution

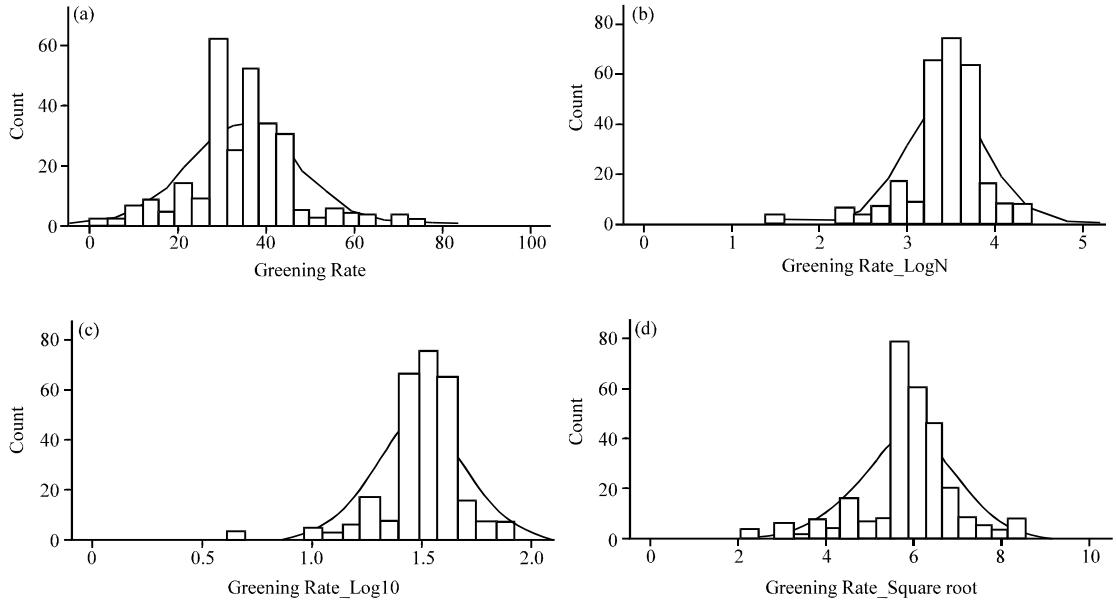

Fig. 4(a-e): Countinue

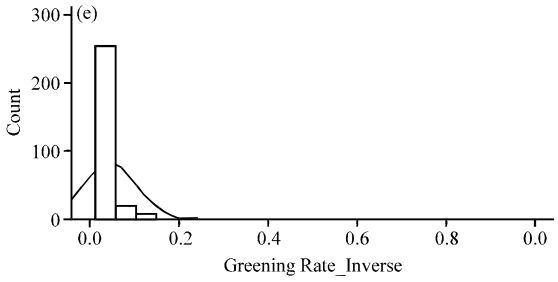

Fig. 4(a-e): Distributions of greening rate (a) Current distribution, (b) lnx distribution, (c) log10×Distribution x1/2, (d) Distribution and (e) x 1 Distribution

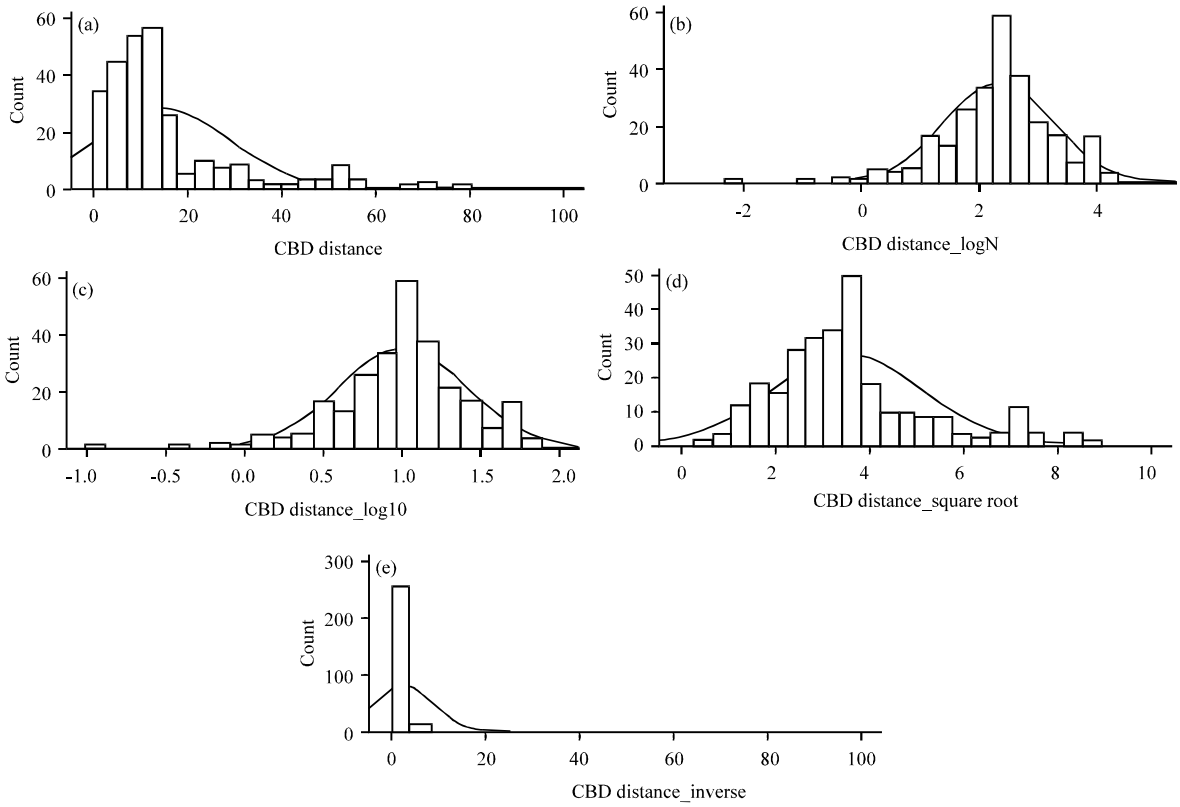

Fig. 5(a-e): Distributions of CBD distance (a) Current distribution, (b) lnx distribution, (c) log10×Distribution x1/2, (d) Distribution and (e) x 1 Distribution

Table14 ANOVA2

Source	SS	df	MS	F
Regression	106.4	20	5.3	57.9
Residual	23. 3	253	0.1	
Total	129.7	273		

Table 15: Comparison of R Square

	Original model	New model
R-square	0.704	0.821
Adjusted R Square	0.681	0.806

R Square has been improved after approximative normalization. R Square and Adjusted R Square in Approximatively normalized Model are greater than that in original model, which means linear correlation between dependent variable and independent variables in Approximatively normalized Model is better than that in original model. Table 15 lists the result.

Table 16: Comparison of R Square

	Original model	New model
linear correlation	0.823	0.835

Testing set is used to test the original model and the approximatively normalized model. The linear correlation

in the approximatively normalized model is better than that in the original model. Table16 lists the result.

CONCLUSIONS

This study analyses whether approximative normalization improves residential hedonic price multivariate linear regression model.

First, selection of the residential characteristics is introduced. Next, Quantification of some numerical residential characteristics variables is introduced. Then, knowledge about multiple linear regression model use in this study is introduced.

In the part of experiment in this paper, two models are established using original data set and data set dealt with by approximative normalization. The two models are compared according to R Square and linear correlation between actual values and predicted values. Finally, it is concluded that approximative normalization can improve residential hedonic price multivariate linear regression model.

ACKNOWLEDGMENTS

This research was supported by Project of National Social Sciences Foundation, Grant No. 13BTJ005,, Social Science Foundation for the Youth Scholars of Ministry of Education of China, Grant No. 10YJC630236, the Fundamental Research Funds for the Central Universities, rant No. 2013XZD01, supported by the Guangdong Province Science and Technology Fund, Grant No. 2012B091100309 and 2012B040500010 and supported by the Foshan Science and Technology Fund, Grant No. 2012HC100043.

REFERENCES

Butler, R.V., 1982. The specification of hedonic indexes for urban housing. Land Econ., 58: 96-108.

Crocker, J., L.L. Thompson, K.M. McGraw and C. Ingerman, 1987. Downward comparison, prejudice and evaluations of others: Effects of self-esteem and threat. J. Personality Soc. Psychol., 52: 907-916.

Dunn, O.J. and V.A. Clark, 1987. Applied Statistics: Analysis of Variance and Regression. John Wiley and Sons Inc., New York.

Guo, W.G., X.M. Cui and H.Z. Wen, 2006. Hedonic price analysis of urban housing: The experiential research on the Hangzhou City. Econ. Geogr., 26: 172-176.

Hao, Q.J. and J. Chen, 2007. Distance to CBD, transport accessibility and geospatial differences of Shanghai residential prices. World Econ. Pap., 1: 22-34.

Liu, J.W., M. Li, J.J. Hou and B. Chen, 2009. The application of linear regression in the problem of real estate business development. http://www.paper.edu.cn/index.php/default/releasepaper/content/34336

Lu, Q.H., 2012. Neural networks in commercial residential pricing-case of Hangzhou. J. Zhejiang Univ. Technol. (Soc. Sci.), 11: 337-341.

Ma, S.X. and A. Li, 2003. House price and its determinations in Beijing based on hedonic model. China Civil Eng. J., 36: 59-64.

Ozanne, L. and S. Malpezzi, 1985. The efficacy of hedonic estimation with the annual housing survey. J. Econ. Soc. Measur., 13: 153-172.

Ren, S., X. Guo and Y.C. Ren, 2009. Multiple linear regression analysis of factors in real estate prices. http://www.paper.edu.cn/releasepaper/content/200907-127

Wang, P., 2008. Analysis of the factors that affect real estate prices. http://www.paper.edu.cn/releasepaper/content/200812-113

Wang, X.Y., 2006. Theoretical analysis of hedonic model of urban housing. Shanghai Manage. Sci., 4: 68-69.

Wen, H.Z. and S.H. Jia, 2006. Market segment and hedonic price analysis of urban housing. J. Zhejiang Univ. Technol., 36: 155-161.

Zhang, M. and S.M. Chen, 2008. Shanghai real estate prices empirical analysis based on hedonic price theory. Financial Econ., 6: 72-74.